An Adaptable Indexing Pipeline for Enriching Meta Information of Datasets from Heterogeneous Repositories

Farshidi, S.; Zhao, Z.

DOI
10.1007/978-3-031-05936-0_37

Publication date
2022

Document Version
Author accepted manuscript

Published in
Advances in Knowledge Discovery and Data Mining

An adaptable indexing pipeline for enriching meta information of datasets from heterogeneous repositories

Siamak Farshidi and Zhiming Zhao*

Multiscale Networked Systems, University of Amsterdam, The Netherlands
{s.farshidi, z.zhao}@uva.nl

Abstract. Dataset repositories publish a significant number of datasets continuously within the context of a variety of domains, such as biodiversity and oceanography. To conduct multidisciplinary research, scientists and practitioners must discover datasets from various disciplines unfamiliar with them. Well-known search engines, such as Google dataset and Mendeley data, try to support researchers with cross-domain dataset discovery based on their contents. However, as datasets typically contain scientific observations or collected data from service providers, their contextual information is limited. Accordingly, effective dataset indexing can be impossible to increase the Findability, Accessibility, Interoperability, and Reusability (FAIRness) based on their contextual information. This paper presents an indexing pipeline to extend contextual information of datasets based on their scientific domains by using topic modeling and a set of suggested rules and domain keywords (such as essential variables in environment science) based on domain experts' suggestions. The pipeline relies on an open ecosystem, where dataset providers publish semantically enhanced metadata on their data repositories. We aggregate, normalize, and reconcile such metadata, providing a dataset search engine that enables research communities to find, access, integrate, and reuse datasets. We evaluated our approach on a manually created gold standard and a user study.

Keywords: dataset indexing · dataset discovery · inverted indexing · metadata standard · data repository.

1 Introduction

Data are increasingly used in decision-making, such as establishing public policies and conducting scientific experiments [17], and are published by various organizations [7], such as scientific publishers, commercial or governmental data providers, research consortia, specialized data repositories, and data aggregators. The more data organizations publish, the more complicated the problem of data discovery becomes [6]. Datasets are typically offered by scientific repositories [1,

* Both authors are corresponding authors. Email: {s.farshidi, z.zhao}@uva.nl

25] or shared via open data portals [29, 19, 21, 35, 28, 14]. Data regarding a set of relevant scientific or practical observations are collected, organized, and formatted for a particular purpose, called dataset [4, 30]. Accordingly, a dataset can be a collection of alphanumeric data, such as entities, diagrams, graphs, design decisions, or textual documents. So that dataset search concerns the discovery, exploration, and retrieval of datasets based on search criteria of searchers [4, 5].

Communities such as Wikidata or the Linked Open Data Cloud [35] offer open and general-purpose data resources that software practitioners can employ in various application domains [7], such as intelligent assistants, recommender systems, and search engine optimization [13, 11, 12]. The primary goal is to increase the findability, accessibility, interoperability, and reusability of zillions of publicly available datasets by enabling data discovery and sharing across organizations within various domains. This trend is reinforced by advances in machine learning and information retrieval, which rely on data to train, validate and enhance their algorithms [32]. To support these applications, we need to search for datasets, which have been researched for decades [8]. However, many characteristics of datasets are unique, with particular requirements and constraints, which have been recognized by well-known dataset search engines, such as Google [6]. According to the literature, we identified the following three challenges in dataset indexing that we are going to address in this study.

$Challenge_1$: General-purpose web search engines typically fail at finding datasets because of *lacking enough description on landing pages of datasets* [16]. In other words, data repositories do not create an individual webpage for each dataset that can be easily recognizable and crawlable by general-purpose web search engines. Data repositories are typically accessible through queries and encrypted web Application Programming Interfaces (Web APIs); this is a well-known phenomenon called *deep Web* [24]. Accordingly, general-purpose search engines index a limited set of datasets.

$Challenge_2$: In literature, various open standards are introduced for describing structured (including dataset metadata) [6]. For instance, Schema.org and the W3C Data Catalog Vocabulary (DCAT) [9] are well-known metadata standards for indexing datasets. Based on our observations (see section 3), and Brickley et al. [6] there is *a limited agreement among dataset repositories in using such metadata*, and they typically define and employ their metadata features to index datasets. Thus, extracting metadata features of datasets from different data repositories automatically based on metadata standards is not possible.

$Challenge_3$: *Links between datasets are still rare, making identifying and using extra contextual information difficult* [6]. In order to offer cross-domain discovery, dataset search engines must improve their ingesting, indexing, and cataloging processes. So that incorporating external knowledge in the data handling process and better management and usage of dataset-intrinsic information can be considered two alternative solutions [7]. Incorporating external contextual information, whether through domain ontologies, tacit knowledge of domain experts, external quality indicators, domain keywords (e.g., essential variables [22]), or even

unstructured information (e.g., in natural language) that describes the datasets, is a fundamental problem.

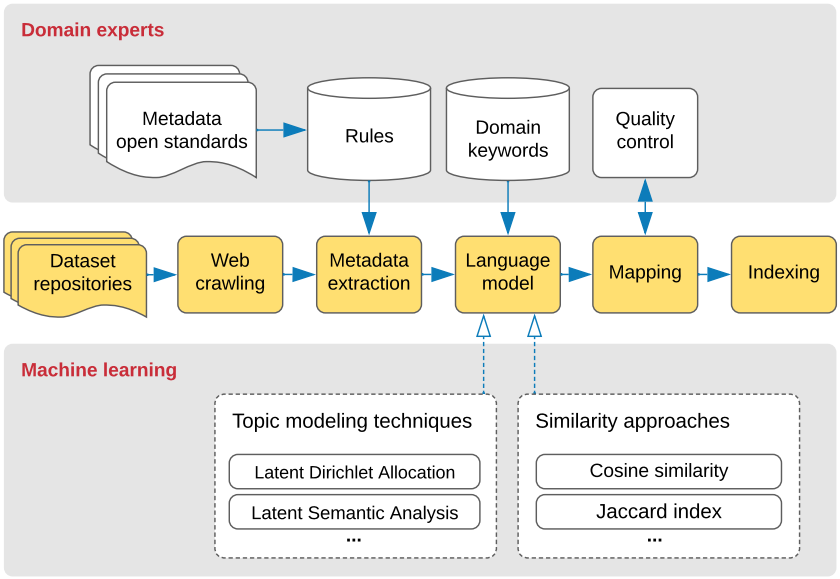
We introduce a novel dataset indexing pipeline to address these three challenges, incorporating information retrieval techniques including web crawling, metadata extraction, language models, human in the loop, and topic modeling to identify semantic similarities and generate indexing documents. The novelty of the proposed pipeline lies in (1) using domain experts' insights to collect *an extendable set of rules* for extracting and refining metadata of dataset records from heterogeneous repositories. Moreover, (2) it employs machine learning techniques, such as topic modeling and similarity approaches, as *replaceable components* to identify topic similarities. Furthermore, (3) the pipeline generates a mapping for each dataset record that adds *additional contextual information* to it. The final mappings can be used to generate effective indexing (e.g., inverted indexing). The proposed pipeline is adjusted based on extendable rules, domain keywords (e.g., essential variables), and domain experts observe and monitor their impacts on the mapping quality.

This paper is structured as follows. Section 2 elaborates on the proposed pipeline and its constituent components. Section 3 explains the experiment that we have conducted with four real-world dataset repositories to evaluate the pipeline. Section 3.4 analyzes the results of the experiment and assesses the performance of the pipeline on the selected dataset repositories. Section 4 discusses the lessons learned, the pipeline limitations, and feedback from the experts.Section 5 concludes this study and highlights our future research directions.

2 Dataset indexing pipeline

In this section, we elaborate on the constituent components of the proposed indexing pipeline. Figure 1 shows the components of the pipeline and its workflow.
Dataset Repositories refer to datasets isolated to be mined for data reporting and analysis. Data repositories are an extensive database of research infrastructures, such as ICOS and SeaDataNet, (see section 3) that collect, manage, and store datasets for data analysis, sharing, and reporting.
Web crawling is the process of a spider bot that systematically browses dataset repositories and extracts dataset records in terms of RDF documents or their landing pages. It retrieves such contents in structured formats (e.g., JSON or key-values). The Web crawling process starts with a list of URLs to visit (seeds). The crawler identifies all the hyperlinks in the retrieved documents/landing pages and adds them to the list of its frontiers to visit them subsequently.
Metadata extraction is the process of retrieving any embedded metadata present in a document. It is responsible for extracting metadata features such as classes and properties inside an RDF document or textual contents of potential features mentioned on landing pages of datasets. The metadata of the retrieved documents will be extracted based on the rules that domain experts define them.
Language model employs various statistical and probabilistic methods to specify the probability of a given sequence of words occurring in a textual document.

Fig. 1. shows the constituent components of the pipeline and its workflow. The pipeline crawls dataset repositories and extracts metadata based on rules that domain experts define according to metadata open standards and domain knowledge. Then, the language model of the pipeline employs topic modeling techniques and similarity approaches to map the domain keywords and extra contextual information to the extracted metadata and create a mapping. The mapping quality will be checked frequently, and the hyperparameters will be adjusted accordingly. Finally, the mapping will be used to create indexes for the records of the dataset repositories.



It analyzes bodies of documents to convert qualitative information into quantitive information. In other words, the language model calculates similarities among metadata features of a particular dataset record and potential contextual information that could be assigned to it. Since contextual information, such as domain keywords, can be seen as vectors, we can use different similarity approaches, such as the cosine similarity or Jaccard index, to calculate the similarity of these vectors [34].

Mapping refers to the process of adding external contextual information, such as domain keywords, to extracted metadata features based on predefined rules by domain experts and language models' predictions. For instance, "sea surface salinity" and "sea surface temperature" as two domain keywords (essential variables[1]) can be mapped to a dataset record that the language model identified the following topics for it: *(water- temperature- dimension- dissolved- salinity- gas- oceanography- chemical- pigment- oceanographic- custodian- sea- geographical- coordinate- spatial)*.

[1] https://earthdata.nasa.gov/learn/backgrounders/essential-variables

Indexing is a data structure technique to efficiently retrieve dataset records on some attributes on which the indexing has been done. Indexing techniques can be used to reduce the processing time of a search query. For instance, inverted indexing categorizes datasets based on collected topics and external contextual information. Then, the final indexes can be ingested in a document data storage (such as ElasticSearch or Apache Solr).

Metadata open standards are high-level documents that establish a common form of structuring and understanding data and include principles and implementation issues for employing the standard. There are many metadata standards purposed for specific disciplines. For instance, Schema.org and DCAT are two metadata open standards that indicate how a dataset should be organized and how it can be related to other types of software assets. In this study, the domain experts suggested Schema.org, DCAT3, and ISO 19115-1:2014 for defining rules that should be employed to index datasets from four real-world dataset repositories (including ICOS, SeaDataNet CDI, SeaDataNet EDMED, and LifeWatch) (see section 3).

Rules are a set of human-made rules, which should be defined by domain experts to increase the accuracy of metadata extraction and refine potential extracted values that can be assigned to the metadata features. An example of potential rules in the rule base is presented as follows. The example shows a metadata feature called "identifier", which is a "unique identifier for this metadata record", and its data type is "PropertyValue/Text". The length of potential values for this metadata feature should be at least 15 characters. The *metadata extraction* component should look for metadata features such as "ISBN", "GTIN", or "UUID" to extract potential values that can be mapped to "identifier".

```
"identifier": [
  "datatype" : "PropertyValue/Text",
  "description": "unique identifier for this metadata record",
  "constraint": ["len(15)"],
  "suggested fields": ["ISBN", "GTIN", "UUID", "URI","URL","id","metadataIdentifier",
                       "gmd:fileIdentifier", "gco:CharacterString","pid"] ], ...
```

Domain keywords are content-related terms that are specific to a particular scientific domain. Terms in glossaries of social studies textbooks or essential variables in environmental sciences are examples of such vocabularies. Domain experts are the main source of knowledge for suggesting domain keywords.

Quality control is an essential phase of the pipeline as the quality of the mapping will be evaluated based on the number of values that mapped correctly to metadata features and the number of potential topics, the number of mapped values to the domain keywords. If the mapping quality is not acceptable, the number of topics of topic modeling algorithms, the threshold of the cosine similarity, and the rules should be revised to improve the mapping quality. This process can be considered as hyperparameter tuning, which is the process of choosing a set of optimal hyperparameters for the similarity approaches, such as cosine similarity and Jaccard index.

Topic modeling techniques are employed in the language model to identify the topics of dataset records. This study uses Latent Dirichlet Allocation (LDA) to find potential topics assigned to datasets. LDA is a generative model for the creation of natural language documents [3]. Note, a topic is a subject discussed in one or more documents. Examples of topics include dataset domains such as "Oceanography" entities such as "SeaDataNet" and long-standing subjects such as "climate change". Each topic is assumed to be represented by a multinomial distribution of words.
Similarity approaches are essential in solving many pattern recognition problems such as classification and clustering. Various similarity approaches, such as Cosine similarity and Jaccard index, are available in the literature to compare two text documents and determine how close their context or meaning are. Various text similarity approaches exist. Typically, similarity approaches have their specification to measure the similarity between two queries. For instance, cosine similarity measures the text-similarity between two documents irrespective of their size in Natural language Processing. The text documents are mainly represented in n-dimensional vector space.

3 Evaluation

One of the well-known issues in evaluating dataset indexing is the lack of benchmarks [7]. So, it is essential to identify a set of appropriate metrics to assess dataset indexing techniques and observe if they mimic information retrieval metrics, such as precision and recall. Such metrics should be employed [20] to evaluate the correctness of the indexing pipeline. In this study, we conducted an experiment in the context of four dataset repositories to assess the pipeline's impact on the quality of mappings and evaluate its effectiveness in addressing the dataset indexing challenges.

3.1 Dataset repositories

The dataset repositories used for the evaluation are based on RDF datasets that have been published by four real-world dataset repositories, namely ICOS, SeaDataNet CDI, SeaDataNet EDMED, and LifeWatch.
(1) ICOS[2] (Integrated Carbon Observation System) is a European-wide greenhouse gas research infrastructure that produces standardized data on greenhouse gas concentrations in the atmosphere and carbon fluxes between the atmosphere, the earth, and oceans. The ICOS dataset repository contains more than *400K* dataset records.
(2) SeaDataNet CDI[3] (Common Data Index service) provides aggregated datasets (collections of all unrestricted SeaDataNet measurements of temperature and salinity by sea basins) and climatologies based on the aggregated

[2] https://data.icos-cp.eu/portal/
[3] https://cdi.seadatanet.org/search

datasets and data from external data sources such as the Coriolis Ocean Dataset for Reanalysis and the World Ocean Database for all the European sea basins and the Global Ocean. The CDI dataset repository contains more than *2,6M* dataset records.
(3) SeaDataNet EDMED[4] covers a wide range of disciplines, including marine meteorology; physical, chemical, and biological oceanography; sedimentology; marine biology and fisheries; environmental quality, coastal and estuarine studies, marine geology, and geophysics. Currently, EDMED contains more than *4K* dataset records, held at over 700 Data Holding Centres across Europe.
(4) LifeWatch[5] provides open data access and facilitates exploratory data analysis of data generated by the local marine-freshwater-terrestrial LifeWatch observatory. The LifeWatch dataset repository contains more than *1,1K* dataset records.

3.2 The pipeline configuration

Rule base - Ten domain experts within geology, oceanography, agriculture, environment, and biology research domains were selected based on their expertise and years of experience to participate in the research and assist us with building the rule base and evaluating the pipeline outcomes. Accordingly, we conducted a survey to identify the features and rules employed to extract metadata from the selected dataset repositories. The experts selected the features we need to use from three metadata standards, including DCAT 3, ISO 19115-1:2014, and Shema.org. It is interesting to highlight that almost less than half of the features that the research infrastructures have been employed in their own metadata were compatible with the metadata open standards [6]
Domain keywords - The domain experts suggested three sets of essential variables [10] based on the domains (atmosphere, oceanography, biodiversity) of the dataset repositories. Note, essential variables are variables known to be critical for observing and monitoring a given facet of the Earth system.
Topic modeling technique - We used Latent Dirichlet Allocation (LDA) as a topic modeling technique for generating potential topics of each dataset record based on its textual explanation.
Similarity approaches - We employed cosine similarity and Jaccard index in this study to estimate similarities among generated potential topics (by the LDA algorithm) of each dataset record and the essential variables. In cosine similarity, data objects in a dataset are treated as a vector. The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It is a measure of similarity for the two sets of data, with a range from *0%* to *100%*. The higher the percentage, the more similar the two populations.

[4] https://edmed.seadatanet.org/search/
[5] https://metadatacatalogue.lifewatch.eu
[6] We published the results of our observations, analysis, script, and contextual information on Mendeley Data [10].

3.3 Experiment

First, we randomly selected 100 datasets from the dataset repositories to generate a training set. Two researchers independently determined the correctness of the mapped domain keywords and potential topics (generated by the LDA algorithms) to the selected datasets. To solve this task, they got the description of those datasets besides their extracted metadata feature, mapped domain keywords, potential topics, and the possibility of taking a deeper look inside the datasets themselves. Finally, we compared their responses, and in the case of inconsistencies, we asked both of them to recheck their responses to reach an agreement between their responses. The training set was used to adjust the hyperparameters, such as the similarity thresholds for both cosine similarity and the Jaccard index, and train the topic model. We altered the hyperparameters dynamically to reach their optimal values for the dataset. Then, we employed the Jaccard index and cosine similarity as the quality control approach to reject irrelevant topics.

One of the main weaknesses of information retrieval measures (including recall, accuracy, and F-measure) is the assumption of binary relevance, with human assessors asked to determine, for a set of documents, which members are relevant to the query and which are not. In other words, human experts are needed to judge the retrieved information and evaluate the effectiveness and efficiency of information retrieval methods [26]. The significant number of dataset records makes it impossible to ask human experts to evaluate the pipeline's outcomes thoroughly. We used a fitness function to assess the quality of the mapping automatically. The fitness function gets the mappings and datasets as its inputs, and then it uses the Jaccard index to assess their relevance. In other words, if a domain keyword is mapped correctly (based on the threshold) to a dataset, it will be highlighted as a true positive. Otherwise, it will be marked as a false positive. We calculated the res of the metrics accordingly.

Table 1 shows the results of the analysis on the dataset repositories (ICOS, SeaDataNet *CDI*, SeaDataNet *EDMED*, and LifeWatch) and 100 randomly selected dataset records form each of them to generate the validation set (contains 400 dataset records). Note, for the sake of validity of our evaluation, the intersection of the training set and the validation dataset records is the empty set.

3.4 Analysis

To evaluate the pipeline components and their impacts on the mapping quality, we perform the experiment incrementally. In each step of the experiment, we evaluate the impact of the absence of each component (Cosine similarity, Rules, Topic mining, and Jaccard index) on the quality of the mappings (involved pipeline components). Note that cosine similarity has been considered the baseline in our analysis to calculate similarity in the language model. Moreover, the fitness function cannot analyze the pipeline's impact on the potential

Table 1. shows the results of the analysis on the dataset repositories (ICOS, SeaDataNet *CDI*, SeaDataNet *EDMED*, and LifeWatch)

Analysis / Dataset repositories		ICOS				CDI				EDMED				LifeWatch			
Involved pipeline components	Cosine similarity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
	Rules	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
	Topic modeling	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
	Jaccard index	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
Domain keywords (essential variables)	Precision (EV)	0.01	0.00	0.01	0.00	0.58	0.58	0.61	1.00	0.48	0.48	0.29	0.66	0.26	0.26	0.09	0.36
	Recall (EV)	0.00	0.00	0.00	0.00	0.21	0.21	0.27	0.23	0.08	0.09	0.15	0.19	0.10	0.10	0.05	0.18
	Accuracy (EV)	0.99	1.00	1.00	1.00	0.59	0.59	0.66	0.63	0.79	0.79	0.72	0.81	0.86	0.86	0.83	0.88
	F (EV)	0.00	0.00	0.00	0.00	0.30	0.30	0.36	0.36	0.13	0.14	0.17	0.28	0.14	0.14	0.06	0.23
Potential topics	Precision (To)	N/A	N/A	0.20	1.00	N/A	N/A	0.41	1.00	N/A	N/A	0.31	1.00	N/A	N/A	0.42	1.00
	Recall (To)	N/A	N/A	0.41	1.00	N/A	N/A	0.71	0.74	N/A	N/A	0.74	0.85	N/A	N/A	1.00	1.00
	Accuracy (To)	N/A	N/A	0.38	1.00	N/A	N/A	0.40	0.78	N/A	N/A	0.29	0.85	N/A	N/A	0.46	1.00
	F (To)	N/A	N/A	0.26	1.00	N/A	N/A	0.51	0.85	N/A	N/A	0.43	0.92	N/A	N/A	0.57	1.00
Mapping	# Mapped values	0.38	0.52	0.53	0.45	0.34	0.42	1.00	0.42	0.48	0.51	0.52	0.42	0.35	0.54	0.55	0.46
Inverted Indexing	# key > 1	0.01	0.01	0.83	0.78	0.51	0.51	0.69	0.74	0.16	0.16	0.48	0.43	0.12	0.12	0.42	0.33
	# keys	101	101	23	18	441	441	320	189	251	251	504	334	858	858	602	401
	# singletone links	100	100	4	4	215	215	99	49	212	212	262	190	755	755	348	268

topics when the topic modeling is not applied. So, in such a scenario, the measures are equal to Not Applicable ("N/A").

The average F-measures of the domain keywords, $F(EV)$, and potential topics, $F(To)$, in Table 1, represent that the pipeline outperforms when all its components are involved.

It has already been shown that LDA does not perform well on short documents in which many different words rarely appear, e.g., messages of short messaging services [36]. It is essential to highlight that the datasets from ICOS typically have limited contextual information, so the $F(EV)$ values have not changed significantly by adding or removing a pipeline component. However, they increase the average F-measures of the rest of the datasets in the validation set. Note, to generate an almost stable list of topics using the LDA algorithm, and we repeated the topic modeling ten times. Increasing the number of iterations leads to higher accuracy and higher time consumption.

The number of assigned values to metadata features (# Mapped values) has been increased by applying the components. As the Jaccard index refines the irrelevant candidate values in the mapping, it reduces the number of mapped values and increases the mapping quality.

Keys are combinations of generated topics and successfully assigned domain keywords. In the last section of Table 1, the quality of the mapping has been evaluated. In the absence of topic modeling and Rules, the performance of the pipeline to generate high-quality mapping and keys decreases significantly. In this scenario, the number of identified keys (# keys) and for generating inverted

indexed have increased. Most of the generated keys were singleton and meaningless quality values. In contrast, when the rules and topic modeling components have been applied, the number of keys decreased as the pipeline rejected low-quality values, and the number of the singleton keys decreased significantly as the pipeline aggregated more keys (# key > 1). For instance, the number of meaningful keys increased from 0.01 (1%) to 0.83 (83%) in the ICOS dataset records by adding topic modeling these two components. It is essential to highlight that applying the Jaccard index can lead to lower numbers of keys, as it further reduces the number of noisy or meaningless keys.

4 Discussion

In the literature, we observed that dataset search had been studied for decades by other researchers and practitioners and can be categorized into two types [7]: general-purpose and domain-specific dataset search. In *general-purpose dataset search approaches* such as Dataverse [1], Elsevier Data Search [25]), open data portals [29, 19, 21, 35, 28, 14] and search engines such as DataMed [33], and Google Dataset Search [27], a collection of public and free datasets, in terms of scientific or practical observations, can be searched through their web portals. These dataset search engines are typically domain-independent, so they are not customized for a particular community. However, *domain-specific dataset search approaches* are designed for searching a set of related observations organized for a particular domain by searchers. This pattern of behavior is particularly marked in data lakes [15, 31], data markets [2, 18], and tabular search [23].

This study identified three challenges that general-purpose and domain-specific dataset search approaches face in their indexing phases. ($Challenge_1$) lack of enough description on landing pages of datasets [16] (deep Web [24]), ($Challenge_2$) a limited agreement among dataset repositories in using metadata standards [6], and ($Challenge_3$) complexity of identifying and using extra contextual information in dataset indexing [7]. To address $Challenge_1$, the pipeline contains an extendable set of domain keywords based on domain experts' insights on the dataset's domain. Domain keywords can improve the findability of dataset records by adding more contextual information to them. The proposed pipeline addresses $Challenge_2$ by suggesting an extendable set of rules based on open standards' definitions and properties. This pipeline component increases the quality of mapping and indexing significantly (see Section 3.4). The pipeline uses topic modeling (e.g., LDA) and similarity approaches to generate potential topics regarding a dataset record according to its contents. Then, it maps the most similar domain keywords to dataset records based on the generated topics.

Probabilistic topic modeling approaches such as LDA employ statistical reasoning to discover underlying patterns of data. As the model hyperparameters should be inferred from observations, the accuracy of statistical reasoning depends on the number of observations. LDA models a dataset as a mixture of topics, and then each word is drawn from one of its topics. Thus, the performance

of LDA can be reduced dramatically in the case of short contextual documents (as happened with ICOS dataset records).

Similarity approaches are low sensitive to semantics. For instance, such methods do not consider the words "marine", "seawater", and "oceanic" as semantically similar. Additionally, they do not distinguish phrases based on their orders and conceptual meaning. For instance, "the ocean color is lighter than sky color" is similar to "the sky color is lighter than the ocean color". It is essential to highlight that similarity approaches, such as the Jaccard index, do not consider word frequency in a given document and count the number of common words in two documents. Accordingly, rare words that are mainly more informative in a document will be ignored. Moreover, the number of repetitions of similar words in two documents would not change the results of such similarity approaches. To sum up, before using a similarity approach in the language model, all its characteristics and behaviors should be investigated. Additionally, a performance testing analysis should be conducted beforehand to select the optimal solution for a particular usage.

Although the pipeline proposed in this study addresses three identified challenges in the literature, there are challenges in the literature regarding FAIRness of dataset discovery that requires profound attention. For instance, European Commission highlighted the following dataset discovery challenges: (1) lack of information that specific datasets exist and are available; (2) a lack of transparency of which public authority maintains datasets; (3) a lack of evidence concerning the terms of reuse; (4) datasets which are made available only in formats that are difficult or expensive to use; (5) complex licensing procedures or restrictive fees; (6) exclusive reuse agreements with one commercial third-party or reused restricted to a government-owned organization.

5 Conclusion and future work

Generating value from data needs the ability to find, access, and make sense of datasets. Many efforts are initiated to support dataset sharing and discovery. For instance, the Google dataset allows users to discover data stored in various online dataset repositories via keyword queries. This study highlighted three challenges that general-purpose and domain-specific dataset search approaches face in their indexing phases. ($Challenge_1$) lack of enough description on landing pages of datasets [16] (deep Web [24]), ($Challenge_2$) a limited agreement among dataset repositories in using metadata standards [6], and ($Challenge_3$) complexity of identifying and using extra contextual information in dataset indexing [7]. To address these challenges effectively, we proposed a novel dataset indexing pipeline based on information retrieval techniques. Next, we conducted an experiment incrementally on the pipeline components to evaluate their effectiveness in addressing the challenges. The results confirmed that the pipeline outperforms when all its components are involved.

Probing deeper, the pipeline presented in this paper also provides a foundation for future work in software asset discovery. We intend to conduct research

to address*software asset recommendation* and *context aware search engines* as our (near) future work.

Acknowledgment

This work has been partially funded by the European Union's Horizon 2020 research and innovation programme, by the project of ARTICONF (825134), ENVRI-FAIR (824068) and BLUECLOUD (862409).

References

1. Altman, M., Castro, E., Crosas, M., Durbin, P., Garnett, A., Whitney, J.: Open journal systems and dataverse integration–helping journals to upgrade data publication for reusable research. Code4Lib Journal **50**(30) (2015)
2. Balazinska, M., Howe, B., Koutris, P., Suciu, D., Upadhyaya, P.: A discussion on pricing relational data. In: In Search of Elegance in the Theory and Practice of Computation, pp. 167–173. Springer (2013)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
4. Borgman, C.L.: The conundrum of sharing research data. Journal of the American Society for Information Science and Technology **63**(6), 1059–1078 (2012)
5. Borgman, C.L.: Big data, little data, no data: Scholarship in the networked world. MIT press (2016)
6. Brickley, D., Burgess, M., Noy, N.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: The World Wide Web Conference. pp. 1365–1375 (2019)
7. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L.D., Kacprzak, E., Groth, P.: Dataset search: a survey. The VLDB Journal **29**(1), 251–272 (2020)
8. Codd, E.F., et al.: Relational completeness of data base sublanguages. IBM Corporation (1972)
9. Data Catalog Vocabulary (DCAT) - Version 3: https://www.w3.org/TR/vocab-dcat-3/, accessed: 2021-09-30
10. Farshidi, S.: The observations, analysis, script, and contextual information regarding this paper. Mendeley Data (2022). https://doi.org/doi: 10.17632/3yb7mhxtyf.1
11. Farshidi, S., Jansen, S.: A decision support system for pattern-driven software architecture. In: European Conference on Software Architecture. pp. 68–81. Springer (2020)
12. Farshidi, S., Jansen, S., Deldar, M.: A decision model for programming language ecosystem selection: Seven industry case studies. Information and Software Technology **139**, 106640 (2021)
13. Farshidi, S., Jansen, S., Fortuin, S.: Model-driven development platform selection: four industry case studies. Software and Systems Modeling **20**(5), 1525–1551 (2021)
14. Find open data: https://data.gov.uk, accessed: 2021-09-30
15. Gao, Y., Huang, S., Parameswaran, A.: Navigating the data lake with datamaran: Automatically extracting structure from log datasets. In: Proceedings of the 2018 International Conference on Management of Data. pp. 943–958 (2018)
16. Goel, S., Broder, A., Gabrilovich, E., Pang, B.: Anatomy of the long tail: ordinary people with extraordinary tastes. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 201–210 (2010)

17. Gohar, M., Muzammal, M., Rahman, A.U.: Smart tss: Defining transportation system behavior using big data analytics in smart cities. Sustainable cities and society **41**, 114–119 (2018)
18. Grubenmann, T., Bernstein, A., Moor, D., Seuken, S.: Financing the web of data with delayed-answer auctions. In: Proceedings of the 2018 World Wide Web Conference. pp. 1033–1042 (2018)
19. Hendler, J., Holm, J., Musialek, C., Thomas, G.: Us government linked open data: semantic. data. gov. IEEE Intelligent Systems **27**(03), 25–31 (2012)
20. Kacprzak, E., Koesten, L., Ibáñez, L.D., Blount, T., Tennison, J., Simperl, E.: Characterising dataset search—an analysis of search logs and data requests. Journal of Web Semantics **55**, 37–55 (2019)
21. Kassen, M.: A promising phenomenon of open data: A case study of the chicago open data project. Government information quarterly **30**(4), 508–513 (2013)
22. Lehmann, A., Masò, J., Nativi, S., Giuliani, G.: Towards integrated essential variables for sustainability (2020)
23. Lehmberg, O., Bizer, C.: Stitching web tables for improving matching quality. Proceedings of the VLDB Endowment **10**(11), 1502–1513 (2017)
24. Madhavan, J., Ko, D., Kot, Ł., Ganapathy, V., Rasmussen, A., Halevy, A.: Google's deep web crawl. Proceedings of the VLDB Endowment **1**(2), 1241–1252 (2008)
25. Mendeley Data: https://data.mendeley.com/research-data/, accessed: 2021-09-30
26. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems (TOIS) **27**(1), 1–27 (2008)
27. Nguyen, T.T., Nguyen, Q.V.H., Weidlich, M., Aberer, K.: Result selection and summarization for web table search. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 231–242. IEEE (2015)
28. Open Data Monitor: https://www.opendatamonitor.eu/, accessed: 2021-09-30
29. Open Knowledge Foundation (CKAN): https://ckan.org/, accessed: 2021-09-30
30. Pasquetto, I.V., Randles, B.M., Borgman, C.L.: On the reuse of scientific data. Data Science Journal **16**, 8 (2017)
31. Reynolds, P., Neuman, K.L., Officer, C.P.: Dhs data framework. dhs.gov (2014)
32. Roh, Y., Heo, G., Whang, S.E.: A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering (2019)
33. Sansone, S.A., Gonzalez-Beltran, A., Rocca-Serra, P., Alter, G., Grethe, J.S., Xu, H., Fore, I.M., Lyle, J., Gururaj, A.E., Chen, X., et al.: Dats, the data tag suite to enable discoverability of datasets. Scientific data **4**(1), 1–8 (2017)
34. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Handbook of latent semantic analysis, pp. 439–460. Psychology Press (2007)
35. The Linked Open Data Cloud: https://www.lod-cloud.net/, accessed: 2021-09-30
36. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European conference on information retrieval. pp. 338–349. Springer (2011)